# A Water Behavior Dataset for an Image-Based Drowning Solution


1st Saifeldin Hasan
*Electrical Engineering Department*
*Rochester Institute of Technology*
Dubai, UAE
skh9588@rit.edu

2nd John Joy
*Electrical Engineering Department*
*Rochester Institute of Technology*
Dubai, UAE
jsj3408@rit.edu

3rd Fardin Ahsan
*Electrical Engineering Department*
*Rochester Institute of Technology*
Dubai, UAE
fxa8225@rit.edu

4th Huzaifa Khambaty
*Electrical Engineering Department*
*Rochester Institute of Technology*
Dubai, UAE
hk7044@rit.edu

5th Manan Agarwal
*ELectrical Engineering Department*
*Rochester Instutute of Technology*
Dubai, UAE
ma7318@rit.edu

6th Jinane Mounsef
*Electrical Engineering Department*
*Rochester Institute of Technology*
Dubai, UAE
jmbcad@rit.edu



2021 IEEE Green Energy and Smart Systems Conference (IGESSC) | 978-1-6654-3456-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGESSC53124.2021.9618700



***Abstract*—Drowning is responsible for an estimated of 320,000 deaths annually worldwide, roughly 25% of those deaths are in swimming pools. This is probably due to the fact that a drowning person, to the untrained eye, will appear to be normally playing or floating in the water. While drowning, a person is unable to call for help, as the nervous system focuses on gathering oxygen for the lungs. To assist the lifeguards with their rescue mission, we propose a water behavior dataset curated to support the design of image-based methods for drowning detection. The dataset includes three major water activity behaviors (swim, drown, idle) that have been captured by overhead and underwater cameras. Moreover, we develop and test two methods to detect and recognize the drowning behavior using the proposed dataset. Both methods use deep learning and aim to support a fast and smart pool rescue system by watching for the early signs of drowning rather than looking for a drowned person. The results show a high performance of the presented methods validating our dataset, which is the first public water behavior dataset and the main contribution of the work.**



***Index Terms*—water behavior dataset, drowning, computer vision, early rescue**


## I. Introduction

The burden of drowning for children has become a leading public health problem [1]. The high rates of drowning are an impediment to achieving reductions of early childhood mortality [2]. Many of these death incidents are attributed to a poor adult supervision. Various drowning scenarios involve newly mobile toddlers who wander off, preschoolers who discover ungated swimming pools, in addition to older children and adults at unguarded public swimming areas, such as beaches, rivers, and residential swimming pools. Effective interventions that mitigate drowning risk will improve health outcomes. Yet, interventions, such as fencing around pools, lifeguards, and flotation devices are not always feasible. A recent study shows a small but alarming number of drowning deaths at public swimming areas that are guarded by professional lifeguards [3]. There is strong evidence that humans simply are not very good at noticing rare events while completing a boring, repetitive task [4], [5]. Although lifeguards are usually highly alert, their task is extremely difficult resulting in egregious examples of inattention. The aquatics industry is acutely aware of the challenges they face to prevent drownings in lifeguarded swimming areas. Addressing the question of how one can help lifeguards complete the highly challenging task of identifying rare events (drownings) while completing a repetitive scanning task is a multifaceted problem that requires a multifaceted solution.

Recently, many research works have been devoted to drowning behavior signs understanding. Lu and Ten [6] presented a vision-based approach to detection of drowning incidents in swimming pools at the earliest possible stage using a number of video clips of simulated drowning. The approach detects, tracks swimmers and parses observation sequences of swimmer features for possible drowning behavioral signs. In [7], a real time drowning detection method uses a HSV thresholding mechanism along with contour detection to detect a drowning person in indoor swimming pools and sends an alarm to the lifeguard if the previously detected person is missing for a specific amount of time. A real-time vision system operating at an outdoor swimming pool is presented in [8]. The system is designed to automatically recognize different swimming activities and to detect occurrence of early drowning incidents. To learn unique traits of different swimming behaviors, the authors simulate and collect unique traits of early drowning behaviors and numerous swimming styles.

The industry has made a shift in the past few years to address the drowning crisis by moving toward more safeguarded pools. This also brought in a generation of drowning prevention products. Many of these products enable the lifeguard with a deeper vision through a 3D monitoring screen [9] or the swimmer with a wearable that tracks how long a swimmer's face is submerged [10]. Other surveillance technologies have advanced to the point of being able to answer a distress call

using artificial intelligence software to monitor the swimmer's activity and detect potential problems when the swimmer has been struggling for more than seven seconds [11].

According to the previous surveys, most of the research has been focused on the detection of the swimmer's location to identify submerged swimmers with no movement. It is not applicable to water activity recognition, in our case such as activities that describe specific behaviors of swimmers. On the other hand, these recognition works applied the model on either simulated swimming scenarios or on their own private real-time video frames. Based on these considerations, this paper first proposes a water activity behavior dataset of videos that are captured above water and under water. The videos illustrate three types of water activities that describe the behaviors of swimming, drowning and staying idle (resting or playing). Moreover, we present two image-based methods that train a deep learning model on the proposed dataset.

The remainder of the paper is organized as follows. Section II presents the two drowning recognition methods. The experimental setup describing the proposed dataset and the results are presented in Section III. Finally, the conclusions are drawn in Section IV.

## II. Drowning Recognition

We utilize existing deep neural networks (DNNs) pre-trained on the large ImageNet dataset [12], and adapt them for water behavior recognition with the sole purpose of identifying the early signs of drowning. The pre-trained feature representations provide a starting point for creating robust classifiers for drowning detection. We consider two scenarios for incorporating pre-trained neural networks. First, we use pre-trained DNNs to re-train them by fine-tuning their parameters. Next, we detect the different body keypoints with a Deep High Resolution Network (HRNet) [13] and use them to train a generic DNN. All the models are trained on a NVIDIA GTX 1070 GPU.

### A. Method 1: Scene Classification

The first method is applied on the proposed dataset to perform a standard video scene classification. The method trains a deep neural network (DNN) using transfer learning in order to perform predictive labelling on the video frames to classify the different human activities in water. We test three different DNN architectures: ResNet50 [14], VGG16 [15], and MobileNet [16]. These networks have been pre-trained on the ImageNet dataset that includes over 1.2 million images for 1,000 object classes. We finally re-train the DNNs by fine-tuning the parameters of the neural networks. Fine-tuning is essentially training the network for several more iterations on a new dataset. This process will adapt the generic filters trained on the ImageNet dataset to the drowning recognition problem. We train the networks for 15 epochs using stochastic gradient descent. At around 15 epochs, all of the network architectures achieve near 100% accuracy on the training set, so no more improvement in training can be achieved.

TABLE I: Summary of the overhead and underwater videos characteristics for each of the three water activities.

| Category | No. of videos (overhead) | No. of frames (overhead) | No. of videos (underwater) | No. of videos (underwater) |
|---|---|---|---|---|
| Swim | 19 | 9,384 | 19 | 8,388 |
| Drown | 18 | 6,080 | 14 | 7,363 |
| Idle | 10 | 9,265 | 11 | 6,259 |

### B. Method 2: Body Pose Estimation

While the features from the pre-trained DNNs can be useful for human activity recognition in the water, another network, the High Resolution Network (HRNet), can be trained to detect and compute the body pose keypoints. A BGR to RGB filter is applied first to the frames before passing them to the DNN to ameliorate the network's performance.The HRNet consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion). The model has been pre-trained on the COCO tain2017 dataset [17] that contains over 50,000 images and 150,000 person instances labeled with 17 keypoints. Finally, the detected keypoints are classified using a DNN of 5 layers with (50,50,50) neurons in the hidden layers to classify the three different human water activities (swim, drown, idle).

## III. Experimental Setup and Results

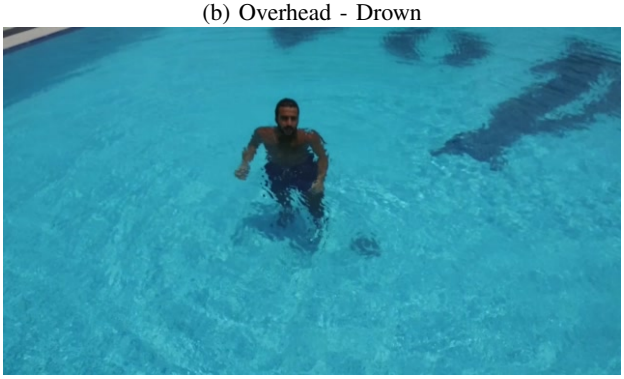
### A. Water Behavior Dataset

To train and test the previously described models, a dataset of short videos displaying water human activities is curated. For this purpose, two different cameras are used above water and under water. The DJI OSMO+ is used to film above water, with a resolution of 1920x1080 at 30 fps, and a GoPRO hero 7 is used for the underwater scenes, with a resolution of 1920x1440 at 30 fps.

For practicality purpose, every video shows only one person performing one of three different activities: swimming, drowning and staying idle. Every video in the dataset is presented and labeled as "activity_id_person.mp4". The underscore ('_') delimiter is used to separate the fields of interest for a better visibility. The first field is the activity type (swim, drown, idle), followed by 'id' to indicate the number of the video and finally 'person' to indicate that we are labelling videos of a person. Individual frames are generated every 0.0333 seconds from their respective videos. The video frames are resized to (640x144) for the scene classification method and to (640x368) for the pose estimation method. The dataset includes a total of 91 videos of an average of 57 minutes each. They are split into overhead and underwater videos, with sub-categories for the different three activities. There are 47 videos for the overhead scenes, and 44 videos for the underwater scenes. The subjects in the dataset are all males ranging from 18 – 21 years of age, mostly of middle eastern ethnicity. Fig. 1 and Fig. 2 show sample images from both overhead and underwater videos for each water activity. Table I presents a summary of the dataset characteristics displaying the number

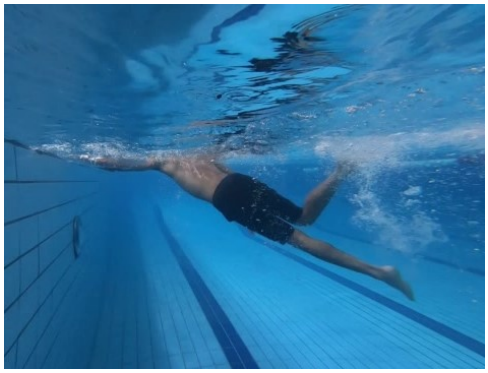
(a) Overhead - Swim

(b) Overhead - Drown

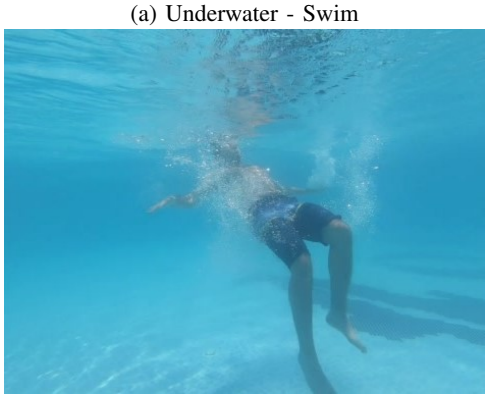


(c) Overhead - Idle

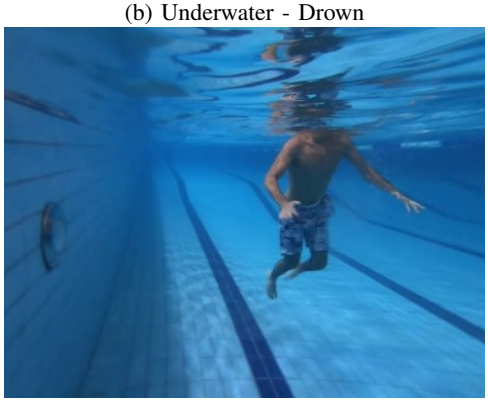
Fig. 1: Sample images from the overhead videos. From top to bottom, images show the swim, drown and idle cases for the same person, respectively.



(a) Underwater - Swim



(b) Underwater - Drown



(c) Underwater - Idle

Fig. 2: Sample images from the underwater videos. From top to bottom, images show the swim, drown and idle cases for the same person, respectively.

of videos and frames in the overhead and underwater scenes for each water activity.

### *B. Experimental Setup*

We split the dataset into training/testing sets for the overhead and underwater cases separately. We select the number of frames in each category (overhead and underwater) to be the same to avoid an unbalanced dataset, where one category (overhead or underwater) is more represented than the other. The training set consists of 90% of the images and the testing set includes the remaining 10%. For each of the overhead and underwater cases, we need to classify the testing set into one of three classes corresponding to swim, drown and idle.

For performance evaluation, we use the accuracy, f1-score, precision and recall recognition rates for both applied methods, scene classification and pose estimation.

### *C. Scene Classification Results*

For the scene classification method, we evaluate the performance of the three DNN architectures. We construct our experiments as described in Section III.B. Table II shows the recognition accuracies of the three models when they are trained on the frames of the three activity classes, as described in Table I. Our results show that ResNet50 significantly

TABLE II: Accuracy (%) of models trained and tested on the proposed water behavior dataset for the three activity classes.

| DNN | Overhead Accuracy (%) | Underwater Accuracy (%) | Average Accuracy (%) |
|---|---|---|---|
| ResNet50 | **98.70** | 95.0 | **96.85** |
| MobileNet | 89.90 | 76.60 | 83.25 |
| VGG16 | 98.30 | **95.10** | 96.70 |

TABLE III: F1-score of models tested on the proposed water behavior dataset for the drowning activity class.

| DNN | Overhead F1-Score | Underwater F1-Score | Average F1-Score |
|---|---|---|---|
| ResNet50 | 0.93 | **0.97** | **0.96** |
| MobileNet | 0.68 | 0.83 | 0.76 |
| VGG16 | **0.95** | **0.97** | **0.96** |

TABLE IV: Performance evaluation of ResNet50 tested on the proposed water behavior dataset for the three activity classes above water.

| Class | F1-Score | Precision | Recall |
|---|---|---|---|
| swim | 0.99 | 0.98 | 0.99 |
| drown | 0.97 | 0.97 | 0.97 |
| idle | 0.98 | 1.00 | 0.99 |

TABLE V: Performance evaluation of ResNet50 tested on the proposed water behavior dataset for the three activity classes under water.

| Class | F1-Score | Precision | Recall |
|---|---|---|---|
| swim | 0.95 | 0.98 | 0.97 |
| drown | 0.93 | 0.94 | 0.94 |
| idle | 0.97 | 0.91 | 0.94 |

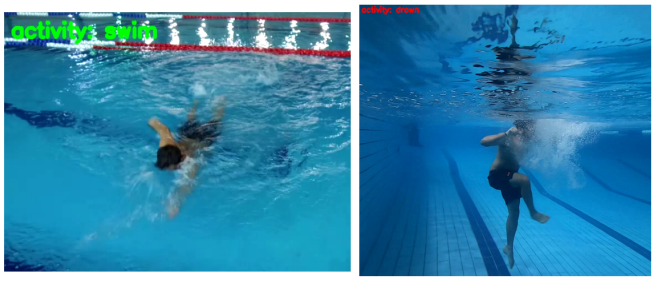
outperforms VGG16 for the overhead and underwater cases, while it performs slightly better than VGG16.

We also analyze the results of our fine-tuning procedure on all three DNN networks for the drowning activity class, in particular, to evaluate the performance of this method at detecting a distress call. Table III shows the f1-score values of the three models when they are tested on the drowning class frames, as described in Table I. The f1-score, in this case, evaluates the false positives and negatives, instead of only assessing the performance of the models at detecting true positives and negatives, as represented by the recognition accuracy. The results show that the ResNet50 and VGG16 significantly outperform the MobileNet network.

Finally, we evaluate the performance of ResNet50 at identifying each of the three different activity classes. Table IV and Table V display the f1-score, precision and recall values to describe the ResNet50 performance for the overhead and underwater cases, respectively. The results show a very good recognition of all three activities. Fig. 3 shows two samples of frames where the recognition outcome is displayed for the swim and drown activity cases using ResNet50.

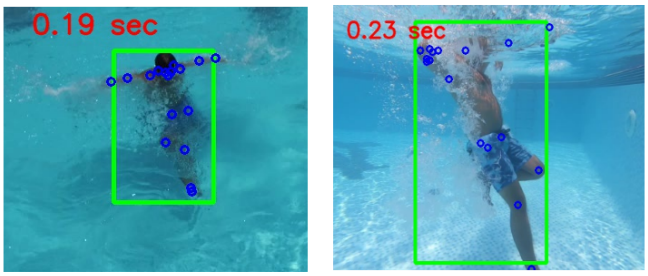
### *D. Pose Estimation Results*

For the pose estimation method, the HRNet detects the pose keypoints in the frames of the testing set, as shown in Fig. 4. The method shows a fast response of less than half a second at detecting the different keypoints across the swimmer's body.




(a) (b)

Fig. 3: Water activity recognition using video scene classification (Method 1) with ResNet50. (a) Swim activity. (b) Drown activity.




(a) (b)

Fig. 4: Water activity recognition using pose estimation classification (Method 2). (a) Above water. (b) Under water.

TABLE VI: Performance evaluation of the pose estimation method tested on the proposed water behavior dataset.

| Category | Accuracy (%) | F1-Score (Drown) |
|---|---|---|
| Overhead | 99.1 | 0.98 |
| Underwater | 97.7 | 0.97 |
| **Average** | **98.4** | **0.98** |

TABLE VII: Performance evaluation of the pose estimation method tested on the proposed water behavior dataset for the three activity classes above water.

| Activity Class | F1-Score | Precision | Recall |
|---|---|---|---|
| swim | 0.99 | 0.99 | 0.99 |
| drown | 0.98 | 0.98 | 0.99 |
| idle | 0.99 | 0.99 | 0.99 |

TABLE VIII: Performance evaluation of the pose estimation method tested on the proposed water behavior dataset for the three activity classes under water.

| Activity Class | F1-Score | Precision | Recall |
|---|---|---|---|
| swim | 0.98 | 0.97 | 0.99 |
| drown | 0.97 | 0.98 | 0.96 |
| idle | 0.98 | 0.98 | 0.98 |

We evaluate the performance of the HRNet architecture in Table VI, which shows the recognition accuracies of the pose estimation method when it is applied to the frames of the three activity classes above water and under water, respectively. Moreover, Table VI shows the f1-score values for the particular case of the drown activity. Our results show that the pose

estimation method performs well for the three activity classes for the overhead and underwater cases, slightly outperforming the scene classification method (Table II and Table III).

Next, we compute the f1-score, precision and recall for each of the three activity classes for both above water and under water, as shown in Table VII and Table VIII. Here too, the performance of the pose estimation method proves to be efficient at recognizing the different water behavior activities. Again, comparing Table VII and Table VIII to Table IV and Table V respectively, we notice that the pose estimation method performs better than the video scene classification method.

We observe that the pose estimation method is independent of the scene variations and solely relies on the body pose and the swimmer's behavior in the water. This is in contrast to the video scene classification method that relies on the features of the overall frame to learn the water activity class. This might be affected by several scene factors that constrain the classification and result in a less robust recognition method compared to the pose estimation method.

## IV. Conclusion

We proposed a water behavior dataset of videos that were captured above water and under water for drowning recognition. The videos illustrated three types of water activities to describe the behaviors of swimming, drowning and staying idle. Moreover, we presented two image-based methods that trained different deep learning models on the proposed dataset. Both methods, the scene classification and pose estimation methods, proved to be efficient at recognizing each of the water behavior activities. In the first method, ResNet50 showed to perform the best. However, the pose estimation method slightly outperformed the scene classification method, knowing that the former depended less on the scene variations. The recognition process relied on keypoint features that described the body pose, which better related to the concept of human behavior in the water.

## Acknowledgment

The authors would like to thank Luke Cunningham and Kim Beasley of BlueGuard - Al Wasl Swimming Academy who supported with creating the proposed water behavior dataset.

## References

[1] D. You, G. Jones, K. Hill, T. Wardlaw and M. Chopra M, "Levels and trends in child mortality, 1990-2009," Lancet, vol. 376, no. 9745, pp. 931–933, September 2010.

[2] M. Peden, K. Oyegbite, J. Ozanne-Smith and A. A. Hyder, "World Report on Child Injury Prevention: Summary," Geneva, Switzerland: World Health Organization, 2008.

[3] Redwoods Group, "Teen dies in accident at YMCA," available at: http://www.redwoodsgroup.com/YMCAs/RiskManagement/ AquaticsAlerts.html. Accessed July 23, 2020.

[4] J. Duncan, G. W. Humphreys GW, "Visual search and stimulus similarity," Psychological Review, vol. 96, pp. 433–458, July 1989.

[5] J. M. Wolfe, T. S. Horowitz, N. M. Kenner, "Rare items often missed in visual searches," Nature, vol. 435, pp. 439–440, May 2005.

[6] W. Lu, Y. P. Tan, "A vision-based approach to early detection of drowning incidents in swimming pools," IEEE transactions on circuits and systems for video technology, vol. 14, no. 2, pp. 159-78, March 2004.

[7] N. Salehi, M. Keyvanara, S. A. Monadjemmi, "An automatic video-based drowning detection system for swimming pools Using active contours," International Journal of Image, Graphics and Signal Processing, vol. 8, no. 8, p. 1, August 2016.

[8] H. L. Eng, K. A. Toh, W. Y. Yau and J. Wang, "DEWS: A live visual surveillance system for early drowning detection at pool," IEEE transactions on circuits and systems for video technology, vol. 18, no. 2, pp. 196-210, Mar 2008.

[9] https://www.aqua-conscience.com/

[10] https://www.wavedds.com/

[11] https://www.angeleye.tech/en/en-lifeguard/

[12] O. Russakovsky, J. Deng, H. Su, J.Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, April 2015.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," arXiv:1902.09212 [cs], Feb. 2019, Accessed: Jun. 15, 2021. [Online]. Available: https://arxiv.org/abs/1902.09212.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, 2012.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, San Diego, May 2015, pp. 1-14.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861. April 2017.

[17] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision, September 2014.